(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 924 687 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
23.06.1999 Bulletin 1999/25

(51) Int Cl.[6]: **G10H 1/00**

(21) Application number: **98309570.4**

(22) Date of filing: **23.11.1998**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **16.12.1997 US 991264**

(71) Applicant: **INTERNATIONAL BUSINESS MACHINES CORPORATION**
Armonk, NY 10504 (US)

(72) Inventors:
- **Vergo, John George**
  Yorktown Heights, NY 10598 (US)
- **Lai, Jennifer Ceil**
  Garrison, New York 10524 (US)

(74) Representative: **Ling, Christopher John**
IBM United Kingdom Limited,
Intellectual Property Department,
Hursley Park
Winchester, Hampshire SO21 2JN (GB)

(54) Speech recognition confidence level display

(57) A speech recognition system and method indicates the level of confidence that a speech recognizer has in its recognition of one or more displayed words. The system and method allow for the rapid identification of speech recognition errors. A plurality of confidence levels of individual recognized words may be visually indicated. Additionally, the system and method allow the user of the system to select threshold levels to determine when the visual indication occurs.

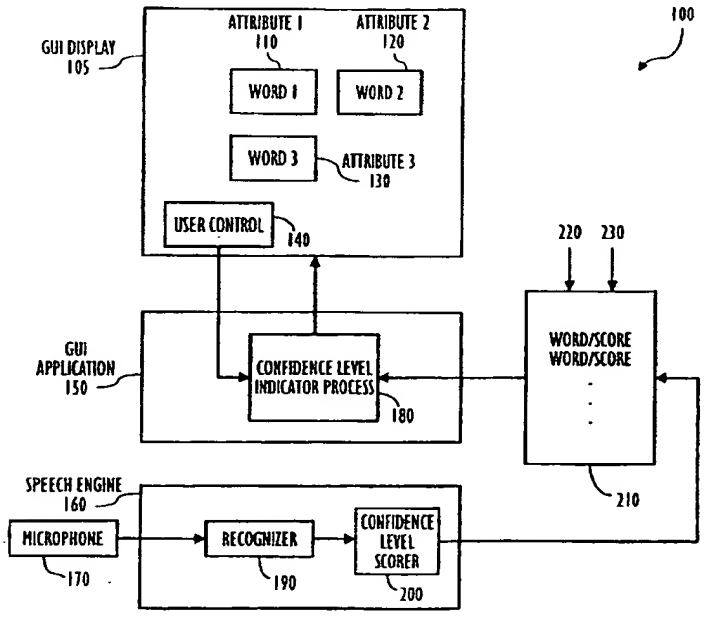

FIG. 1

EP 0 924 687 A2

Description

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] This invention relates to the field of speech recognition systems. More specifically, this invention relates to user interfaces for speech recognition systems, and yet more specifically to a method and apparatus for assisting a user in reviewing transcription results from a speech recognition dictation system.

Description of the Related Art

[0002] Text processing systems, e.g. word processors with spell checkers, such as Lotus WordPro™ and Word Perfect™ by Novell, can display misspelled words (i.e. words not recognized by a dictionary internal to the word processor) in a colour different from that of the normal text. As a variant, Microsoft Word™ underlines misspelled words in a colour different from that of the normal text. In these cases, it is simple to ascertain the validity of a word by checking it against dictionaries. Either a word is correctly spelled or it is not. However, these aspects of known text processing systems deal only with possible spelling errors. Additionally, because spell-checkers in text processing systems use only a binary, true/false criterion to determine whether a word is correctly (or possibly incorrectly) spelled, these systems will choose one of two colours in which to display the word. In other words, there are no shades of gray. The word is merely displayed in one colour if it is correctly spelled and in a second colour if the system suspects the word is incorrectly spelled. Grammar checking systems operate similarly, in that the system will choose one of two colours in which to display the text depending upon whether the system determines that correct grammar has been used.

[0003] By contrast, the inventive method and apparatus of the present invention deals with speech recognition errors, and in particular with levels of confidence that a speech recognition system has in recognizing words that are spoken by a user. With the method and apparatus of the present invention, an indication is produced, which is correlated to a speech recognition engine's calculated probability that it has correctly recognized a word. Whether or not a word has been correctly recognized, the displayed word will always be correctly spelled. Additionally, the inventive system supports multiple levels of criteria in determining how to display a word by providing a multilevel confidence display.

[0004] In another area, known data visualization systems use colour and other visual attributes to communicate quantitative information. For example, an electroencephalograph (EEG) system may display a colour contour map of the brain, where colour is an indication of amplitude of electrical activity. Additionally, meteorological systems display maps where rainfall amounts or temperatures may be indicated by different colours. Contour maps display altitudes and depths in corresponding ranges of colours. However, such data visualization systems have not been applied to text, or more specifically, to text created by a speech recognition/dictation system.

[0005] In yet another area, several speech recognition dictation systems have the capability of recognizing a spoken command. For example, a person dictating text, may dictate commands, such as "Underline this section of text", or "Print this document". In these cases, when the match between the incoming acoustic signal and the decoded text has a low confidence score, the spoken command is flagged as being unrecognized. In such a circumstance, the system will display an indication over the user interface, e.g. a question mark or some comment such as "Pardon Me?". However, obviously such systems merely indicate whether a spoken command is recognized and are, therefore, binary, rather than multilevel, in nature. In the example just given, the system indicates that it is unable to carry out the user's command. Thus, the user must take some action. Such systems fail to deal with the issue of displaying text in a manner that reflects the system's varying level of confidence in its ability to comply with a command.

[0006] In yet another area, J.R. Rhyne and G.C. Wolf's chapter entitled "Recognition Based User Interfaces," published in Advances in Human-Computer Interaction, 4:216-218, Ablex, 1993, R. Hartson and D. Hix, editors, states "the interface may highlight the result just when the resemblance between the recognition alternatives are close and the probability of a substitution error is high." However, this is just another instance of using binary criteria and is to be contrasted with the multilevel confidence display of the present invention. Furthermore, this reference merely deals with substitution error and lacks user control, unlike the present invention which addresses not only substitution errors but also deletion errors, insertion errors, and additionally, provides for user control.

[0007] Traditionally, when users dictate text using speech recognition technology, recognition errors are hard to detect. The user typically has to read the entire dictated document carefully word by word, looking for insertions, deletions and substitutions. For example, the sentence "there are no signs of cancer" can become "there are signs of cancer" through a deletion error. This type of error can be easy to miss when quickly proof reading a document.

[0008] It would be desirable to provide a system that displays transcribed text in accordance with the system's level of confidence that the transcription is accurate. It also would be desirable if such a system could display more than a binary indication of its level of confidence.

DISCLOSURE OF THE INVENTION

[0009] The present invention relates to a speech recognition computer system and method that indicates the level of confidence that a speech recognizer has in one or more displayed words. The level of confidence is indicated using an indicator, such as colour, associated with the word or words that are displayed on a user interface. The system has a voice input device, such as a microphone, that inputs acoustic signals to the speech recognizer. The speech recognizer translates the acoustic signal from the voice input device into text, e. g. one or more words. A confidence level process in the speech recognizer produces a score (confidence level) for each word that is recognized. A confidence level indicator process then produces one, of one or more indications, associated with each of the one or more words displayed on the user interface. The indication is related to one of one or more sub-ranges, in which the score falls. The words are displayed on a user interface as text with the properties of the text (e.g. colour) reflecting the confidence score.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] The invention will now be described, by way of example only, with reference to the accompanying drawings, in which:

Figure 1 is a block diagram of a preferred embodiment of the present invention;

Figure 2 is a flow chart which shows the steps carried out in the system depicted in Figure 1; and

Figure 3 is a flow chart which provides greater detail of the confidence level indicator process.

DETAILED DESCRIPTION OF THE INVENTION

[0011] Figure 1 shows a system and method for displaying words with attributes that are correlated to confidence levels. A human speaker talks into a microphone (170). The microphone transmits an acoustic (speech) signal to a speech engine process (160). The speech engine process may be either software or a combination of software and hardware, which digitizes the incoming acoustic signal and performs a recognition function (190). The recognition function (190) translates the acoustic signal into text, i.e. one or more words. This recognition and translation may be accomplished in a number of different ways which are well known to those in the field. Each word is assigned a confidence level score by a confidence level scorer (200). This confidence level score is assigned using an algorithm to determine the level of accuracy with which the recognizer (190) determines it has translated the acoustic (speech) signal to text. Each word and its assigned confidence level score form a word/score (210) pair, each of which is sent to a graphical user interface (GUI) application (150). The GUI application (150) may receive information from a user control (140) to enable a user of the system to select score thresholds, above which (or below which) default attributes are used in displaying the words. The user may also provide information, via the use control (140), to control which colour maps and/or attribute maps are used to display the words. The use of the thresholds and maps will be discussed in more detail below.

[0012] Having received the word/score pairs, GUI application (150) uses a Confidence Level Indicator Process (CLIP) (180) along with information from the user control (140), if any, to assign a colour and/or an attribute to each word (110, 120, 130). The CLIP is a mapping algorithm which takes the score which was assigned by the confidence level scorer (200) and determines what colour and/or attribute should be associated with that score. The resulting colour and/or attribute used to display the word then reflects the level of accuracy with which the recognizer determines it has translated the acoustic (speech) signal into text.

[0013] The colour selected might be from a map of a range of different colours or might be from a map of different shades of a single colour. Additionally, the attribute selected may include features such as font type, point size, bold, italics, underline, double underline, capitalization, flashing, blinking, or a combination of any of these features. Once a word and its associated colour and/or attribute are determined for each word, the pairs are then displayed on an output device (105), with each word being displayed with its associated colour and/or attribute (110, 120, 130).

[0014] Figure 2 shows, in a flow chart form, the steps which are carried out in the embodiment described in connection with Figure 1. Figure 2 shows that the acoustic (speech) signal generated by a speaker speaking into a microphone is sent to the speech engine process (160) containing a recognizer (190) for decoding the acoustic signal to text or words as well as a confidence level scorer (200) for assigning a score to the words. This score reflects the level of confidence the speech recognition system has in its translation of the processed acoustic signals. Each word, with its associated score is then sent from the confidence level scorer (200) in the speech engine process (160) to graphical user application (150). The graphical user application (150) may accept information from the user control (140) to control the threshold and colour and/or attribute mapping and use that information in the CLIP (180) within the graphical user application (150). The CLIP (180) then assigns a colour and/or attribute to each word based upon the score given to each word and based upon the information from the user, if any. Thus, the graphical user interface application (150) has as its output each word with an associated colour and/or attribute. This information is then used to display the word

with the associated colour and/or attribute, which, in turn, is an indication of the confidence level associated with each word.

[0015] Figure 3 depicts a flow chart showing more detail of CLIP (180 in Figures 1 and 2). A word/score pair (210) is received by the CLIP (180) which assigns a default colour and font attribute to the word (181). The word and its score are reviewed (182). If the word is above the threshold it is displayed with the default colour and attribute (220). If the score is below the threshold (141), which may be defined by a user or defined by the system, the word and its associated score go to a process that checks for colour mapping (183). When a colour map (240) is used, the appropriate colour (determined by the word's score) is mapped to the word (185). Irrespective of whether colour mapping is used, the process checks whether the attribute mapping of the word needs to be changed based on the score (184). If so, the attribute mapping process (184) maps the correct font attribute based on the score (186) using an attribute map (230). The word, with colour and attribute if appropriate, then are displayed (220).

[0016] Variants to the invention are possible. For example, in the flow chart of Figure 3, colour and/or attribute mapping may be carried out if the word/score pair is above, rather than below a threshold. Also, colour mapping or attribute mapping may be carried out alone, rather than serially. That is, either colour mapping or attribute mapping may be used alone.

Claims

1. A speech recognition system comprising:

   a speech recognizer for translating speech into text, said text being one or more words, said speech recognizer further comprising a confidence level scorer (200) for assigning one of at least three possible scores for each of said one or more words, said score being a confidence measure that said one or more words has been recognized correctly; and
   a user interface (150) for displaying said one or more words, each of said one or more words having display properties based on said scores.

2. A speech recognition system as claimed in claim 1, wherein said different display properties include a default display property and two or more other display properties.

3. A speech recognition system as claimed in claim 2, wherein said default display property is normal text.

4. A speech recognition system as claimed in claim 2, wherein said one or more words is displayed with one of said two or more other display properties when said confidence measure is below a threshold, thereby indicating a possible error.

5. A speech recognition system as claimed in claim 4, wherein said threshold level is selected by a user of said speech recognition system.

6. A speech recognition system as claimed in claim 2, wherein said one or more words are displayed with said default display property when said confidence measure is above a threshold level.

7. A speech recognition system as claimed in claim 1, wherein each of said different display properties is a different colour.

8. A speech recognition system as claimed in claim 1, wherein each of said different display properties is at least one different font attribute selected from the group consisting of font type, point size, bold, italics, underline, double underline, capitalization, flashing and blinking.

9. A speech recognition system as claimed in claim 1, wherein each of different display properties is one of a different shade of a single or a different shade of gray.

10. A speech recognition system as claimed in claim 5, wherein said threshold selection enables said user to select one of a colour map or a gray scale map to identify which one of said at least three possible scores is assigned to each of said one or more words.

11. A method of speech recognition comprising:

    translating input speech into text, said text being one or more words;
    assigning one of at least three possible confidence level scores for each of said one or more words, said score being a confidence measure that said one or more words has been recognized correctly; and
    displaying said one or more words based on said assigning step, each of said one or more words having display properties based on said scores.

12. A method of speech recognition as claimed in claim 11, wherein said one or more words is displayed with one of said two or more other display properties when said confidence measure of said one or more words is below a threshold level.

13. A method of speech recognition as claimed in claim 12, further comprising the step of:
    providing user selectability of said threshold

level.

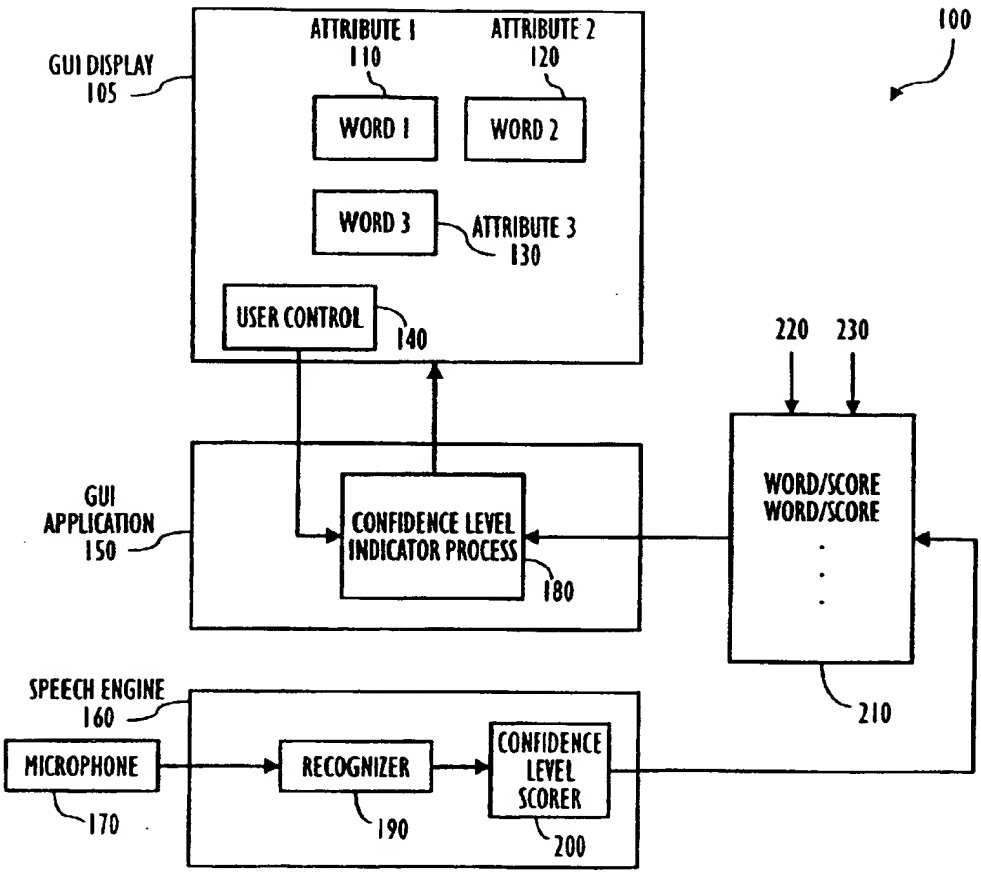
100
GUI DISPLAY
105
ATTRIBUTE 1
110
ATTRIBUTE 2
120
WORD 1
WORD 2
WORD 3
ATTRIBUTE 3
130
USER CONTROL
140
220
230
GUI
APPLICATION
150
CONFIDENCE LEVEL
INDICATOR PROCESS
180
WORD/SCORE
WORD/SCORE
210
SPEECH ENGINE
160
MICROPHONE
170
RECOGNIZER
190
CONFIDENCE
LEVEL
SCORER
200

FIG. 1

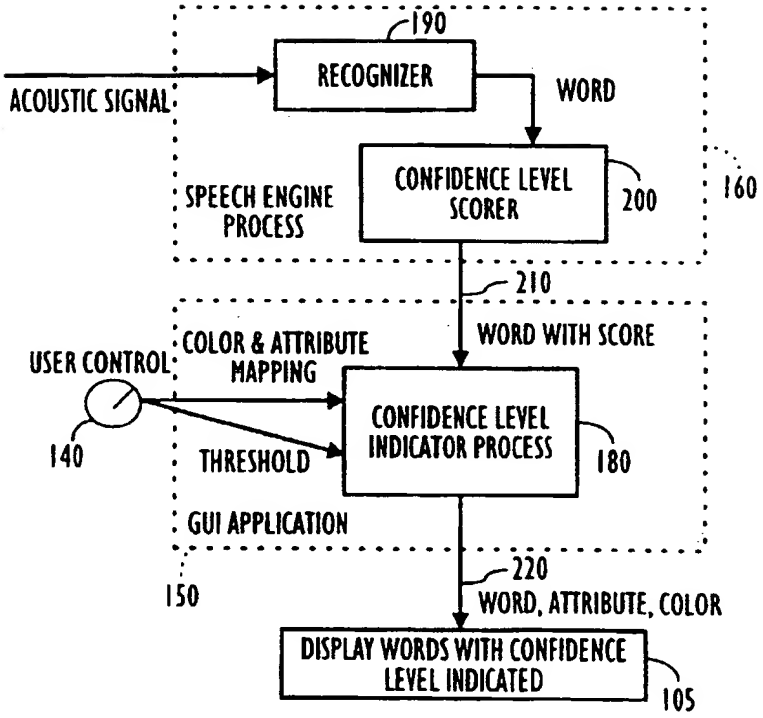
190
RECOGNIZER
ACOUSTIC SIGNAL
WORD
SPEECH ENGINE PROCESS
CONFIDENCE LEVEL SCORER
200
160
210
WORD WITH SCORE
COLOR & ATTRIBUTE MAPPING
USER CONTROL
140
THRESHOLD
CONFIDENCE LEVEL INDICATOR PROCESS
180
GUI APPLICATION
150
220
WORD, ATTRIBUTE, COLOR
DISPLAY WORDS WITH CONFIDENCE LEVEL INDICATED
105

FIG. 2

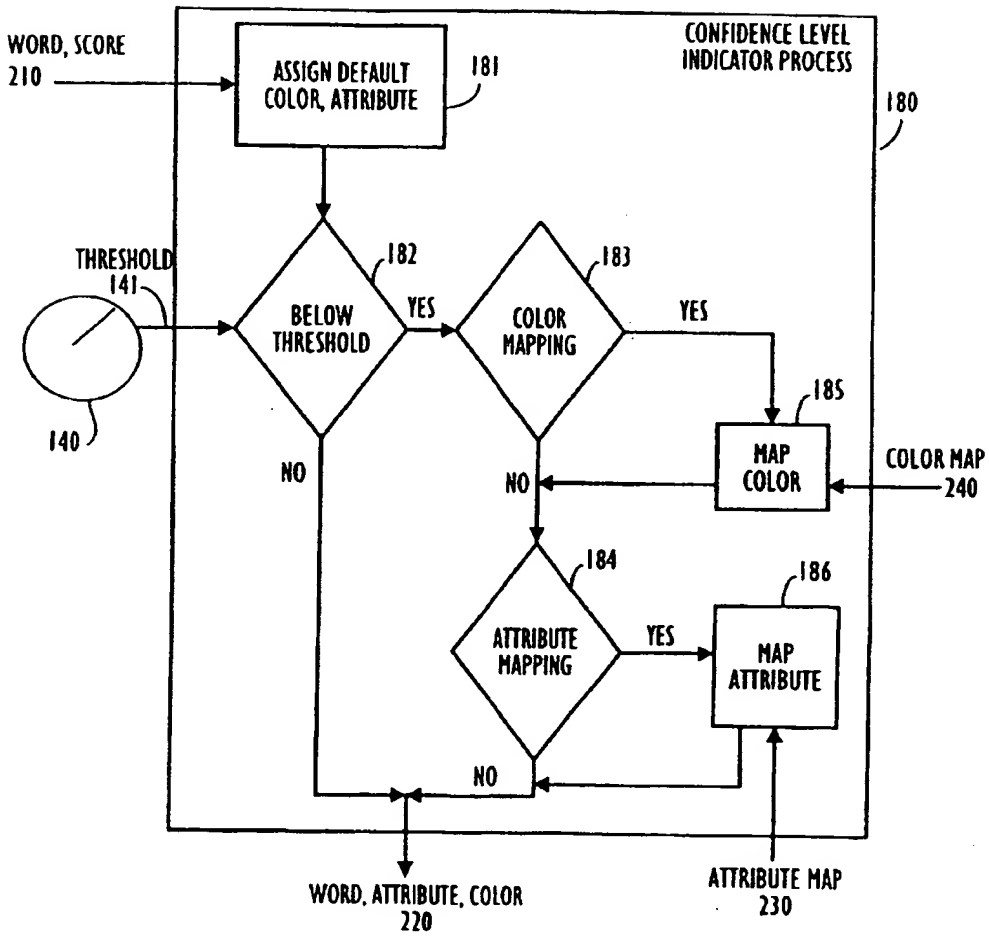
WORD, SCORE
210
ASSIGN DEFAULT COLOR, ATTRIBUTE
181
CONFIDENCE LEVEL INDICATOR PROCESS
180
THRESHOLD
141
140
BELOW THRESHOLD
182
YES
COLOR MAPPING
183
YES
185
MAP COLOR
COLOR MAP
240
NO
NO
184
ATTRIBUTE MAPPING
YES
186
MAP ATTRIBUTE
NO
WORD, ATTRIBUTE, COLOR
220
ATTRIBUTE MAP
230

FIG. 3